How B-cell receptor repertoire sequencing can be enriched with structural antibody data

Kovaltsuk, Aleksandr ; Krawczyk, Konrad ; Galson, Jacob D ; Kelly, Dominic F ; Deane, Charlotte M ; Trück, Johannes

Abstract: Next-generation sequencing of immunoglobulin gene repertoires (Ig-seq) allows the investigation of large-scale antibody dynamics at a sequence level. However, structural information, a crucial descriptor of antibody binding capability, is not collected in Ig-seq protocols. Developing systematic relationships between the antibody sequence information gathered from Ig-seq and low-throughput techniques such as X-ray crystallography could radically improve our understanding of antibodies. The mapping of Ig-seq datasets to known antibody structures can indicate structurally, and perhaps functionally, uncharted areas. Furthermore, contrasting naïve and antigenically challenged datasets using structural antibody descriptors should provide insights into antibody maturation. As the number of antibody structures steadily increases and more and more Ig-seq datasets become available, the opportunities that arise from combining the two types of information increase as well. Here, we review how these data types enrich one another and show potential for advancing our knowledge of the immune system and improving antibody engineering.

DOI: https://doi.org/10.3389/fimmu.2017.01753

frontiers in Immunology

REVIEW
published: 08 December 2017
doi: 10.3389/fimmu.2017.01753

How B-Cell Receptor Repertoire Sequencing Can Be Enriched with Structural Antibody Data

Aleksandr Kovaltsuk[1], Konrad Krawczyk[1], Jacob D. Galson[2], Dominic F. Kelly[3], Charlotte M. Deane[1]† and Johannes Trück[2]*†*

[1] Department of Statistics, University of Oxford, Oxford, United Kingdom, [2] Division of Immunology and the Children's Research Center, University Children's Hospital, University of Zürich, Zürich, Switzerland, [3] Oxford Vaccine Group, Department of Paediatrics, University of Oxford and the NIHR Oxford Biomedical Research Center, Oxford, United Kingdom

OPEN ACCESS

Edited by:
Gregory C. Ippolito, University of Texas at Austin, United States

Reviewed by:
Christopher Vollmers, University of California, Santa Cruz, United States
Jeffrey J. Gray, Johns Hopkins University, United States
Jeliazko R. Jeliazkov, Johns Hopkins University, United States

***Correspondence:**
Charlotte M. Deane
deane@stats.ox.ac.uk;
Johannes Trück
johannes.trueck@kispi.uzh.ch

†Joint senior authors.

Specialty section:
This article was submitted to B Cell Biology, a section of the journal Frontiers in Immunology

Received: *11 October 2017*
Accepted: *27 November 2017*
Published: *08 December 2017*

Citation:
Kovaltsuk A, Krawczyk K, Galson JD, Kelly DF, Deane CM and Trück J (2017) How B-Cell Receptor Repertoire Sequencing Can Be Enriched with Structural Antibody Data. Front. Immunol. 8:1753. doi: 10.3389/fimmu.2017.01753

Next-generation sequencing of immunoglobulin gene repertoires (Ig-seq) allows the investigation of large-scale antibody dynamics at a sequence level. However, structural information, a crucial descriptor of antibody binding capability, is not collected in Ig-seq protocols. Developing systematic relationships between the antibody sequence information gathered from Ig-seq and low-throughput techniques such as X-ray crystallography could radically improve our understanding of antibodies. The mapping of Ig-seq datasets to known antibody structures can indicate structurally, and perhaps functionally, uncharted areas. Furthermore, contrasting naïve and antigenically challenged datasets using structural antibody descriptors should provide insights into antibody maturation. As the number of antibody structures steadily increases and more and more Ig-seq datasets become available, the opportunities that arise from combining the two types of information increase as well. Here, we review how these data types enrich one another and show potential for advancing our knowledge of the immune system and improving antibody engineering.

Keywords: Ig-seq, antibody modeling, B cell, Antibodies, Developability, computational modeling, Next-generation sequencing

INTRODUCTION

Antibodies are proteins produced by the B cells of jawed vertebrates. Their primary function is to recognize structural sequence motifs (epitopes) within molecules (antigens) usually related to pathogens, which may lead to direct neutralization of those pathogens or their toxins. Further functions of antibodies are activation of the complement system or tagging of antigens for elimination by other immune pathways. Antibodies have the capacity for binding an extraordinary variety of epitopes as a result of their sequence diversity, which is estimated at 10^{13} unique molecules in the human antibody repertoire (1). An antibody is a large complex molecule (~150 kDa). It can be divided into two parts, the crystallizable fragment (Fc) and the antigen binding fragment (Fab). The Fab fragment is further split into constant and variable regions. There are five possible main Fc portions in humans, and which one is used on a particular antibody is governed by the process of class switching (2). The variable region (Fv) is composed of two domains called the heavy (VH) and light (VL) chains. Within each B cell, the antibody Fv domains are built by somatic recombination between V(D)J segments (3, 4). Upon antigen recognition, somatic hypermutation introduces further diversification into the naïve Fv domains (5). Within each of the VL and VH chains lie three hypervariable loops, the complementarity determining regions (CDRs), which are the most

diverse parts of the antibody (**Figure 1**). These loops form the majority of chemical interactions with antigens, thus defining the antigen-binding region, the paratope (6). The CDR3 of the heavy chain (CDR-H3) is the most diverse of the CDRs as it is being formed at the join between the V, D, and J gene segments and subject to high levels of hypermutation. As a result of this diversity, CDR-H3 plays a key role in antigen recognition and binding (7). The non-CDR sections of the variable domain are called the framework. Framework positions next to CDRs along with CDR sequence govern the structural shape of the loops (8).

The properties of antibodies, in particular designable antigen recognition specificity and binding affinity, have made them useful as diagnostics and research agents as well as the most successful class of biopharmaceuticals (10). Although small molecules constitute the largest proportion of potential therapeutics in clinical trials, the antibody market is steadily growing, with new antibody approvals at a rate of about four per year. As of 2016, five out of the 10 best-selling drugs worldwide were recombinant monoclonal antibodies (11).

Successful exploitation of antibodies relies on our ability to interrogate their diversity and function. Application of next-generation sequencing of immunoglobulin gene repertoire (Ig-seq) to antibody profiling is able to produce comprehensive snapshots of the repertoire diversity (12). However, most Ig-seq techniques are currently unable to perform sequencing of paired heavy–light antibody sequences or to obtain an immunoglobulin gene repertoire solely from antibody-secreting B cells (13–15). Advances in liquid chromatography tandem-mass spectroscopy (LC-MS/MS) now allow high-throughput analysis of serum antibodies at the amino-acid sequence level (16, 17). Previously transcriptomics and Ig-seq datasets have been used to deconvolute MS spectra of serum antibodies into constituent full-length entities (18). Such combined Ig-seq and LC-MS/MS techniques have provided new insights in vaccination and autoimmunity studies (19, 20). Recent advances in computational tools that integrate *de novo* antibody sequencing, error correction data, and sequence homology databases now permit an accurate assembly of full-length antibodies based on the remit of LC-MS/MS spectra alone (21).

The biggest advantage of Ig-seq and LC-MS/MS techniques is their high-throughput nature. This means that the methods provide a broad-brush description and quantification of antibodies in the repertoire. However, this will often include inaccurate data caused by PCR or sequencing errors. The limitation of Ig-seq and LC-MS/MS methods is that they provide sequence information only, whereas it is the shape/structure of an antibody that determines its exact biological function. For instance, antibody CDRs with low-sequence identities can adopt structurally close shapes,

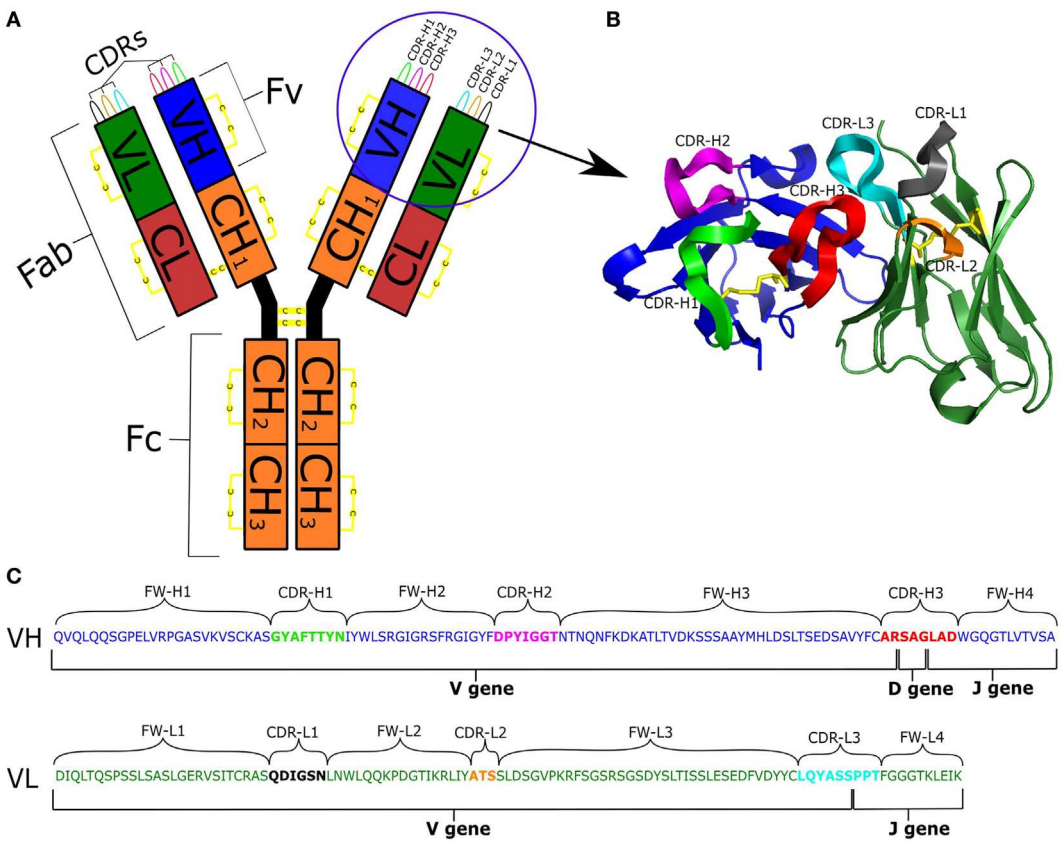


FIGURE 1 | **(A)** Schematic representation of an antibody IgG structure. **(B)** Structure of the Fv region. **(C)** Genetic composition of VH and VL chains [IMGT numbering (9)]: VH is colored blue; VL is green; CDRs are labeled and depicted in different colors; and disulfide bonds are in yellow.

and hence present conformationally similar, though perhaps chemically different, binding sites (22). Knowledge of antibody structure is vital for inferring chemistry of antigen recognition as well as allowing binding site comparison between antibodies. Current experimental determination of antibody structures is achieved by X-ray crystallography or NMR spectroscopy. However, collecting such detailed experimental information limits the rate of analysis to the level of individual or a small number of antibodies (23).

To help tackle the rising costs and time required for engineering and characterization of antibodies, a number of computational tools have been developed that can facilitate experimental efforts. Computational methods are used to profile the physico-chemical properties of antibodies, predict antibody–antigen contacts, and redesign antibody–antigen complexes (24, 25). The tools can be broadly divided into those that require only the sequence of an antibody as input and those that require the structure of the antibody. The inclusion of structural information where available has been shown to improve prediction of most properties over sequence-based methods (26). These improved predictions are only possible if a native structure or an accurate model of the antibody is available.

Since the structure of an antibody is key to its function and high-throughput crystallographic determination of the structures of every antibody is currently not feasible, computational modeling techniques may aid to reduce attrition in the biopharmaceutical industry and to accelerate drug discovery (27). The development of systematic relationships between the antibody information gathered from Ig-seq and techniques such as X-ray crystallography, NMR spectroscopy, and tandem LC-MS/MS could radically improve our understanding of antibody biology. As the number of antibody structures steadily increases and more Ig-seq datasets become available, the opportunities that arise from combining them increase as well. As of October 9, 2017, more than 2,860 antibody structures were available in the Protein Data Bank (PDB) (28) as identified by the Structural Antibody Database (29). The publically available volume of sequences produced from Ig-seq experiments is now in the hundreds of millions (30). In this manuscript, we consider the information obtained from high-throughput sequencing experiments and antibody structures. We review how these datasets can enrich one another and with the help of computational techniques, advance our knowledge of antibody diversity, maturation, and selection and pave the way for improved antibody engineering.

IMMUNOGLOBULIN GENE REPERTOIRE SEQUENCING TECHNOLOGIES

Ig-seq offers high-throughput characterization of immunoglobulin gene sequences at great depth and typically includes several B-cell samples in a single-sequencing run. By controlling the number of samples that are combined and the number of B cells contained therein, it is possible to obtain a large fraction of an immunoglobulin repertoire from a sample. The potential applications of Ig-seq include vaccine and drug development as well as immunodiagnostics (12, 31, 32). Such applications rely on our ability to efficiently identify the population of antibodies responding to an antigen challenge. Ig-seq has already been successfully applied to isolate antigen-specific antibodies from immunized animals in conjunction with common laboratory screening platforms such as phage display (33) or hybridoma (34) or even when the screening step was omitted (35). Furthermore, amino-acid sequence convergences in the CDR-H3 have been observed in the response to a variety of antigens, and may serve as an additional way to isolate antigen-specific antibodies through identifying sequences common among several individuals exposed to the same antigen (30, 36–39).

Heavy and light chains are products of two independent mRNA transcripts that co-assemble into full-length immunoglobulin molecules in the endoplasmic reticulum of the B cell. However, cognate pairing is lost after B-cell bulk lysis prior to Ig-seq and most Ig-seq studies therefore only consider heavy chains (12). However, for human and mouse native pairing is crucial for antibody folding, stability, expression, and antigen binding (40–42). Furthermore, information on the heavy/light chain dimer is required to create an accurate three-dimensional (3D) model of the Fv region and of its antigen-binding pocket which is essential for rational antibody engineering (43). Such models can map antibody sequences to structural space (44), identify the paratope and its physico-chemical properties (45), interrogate the mode of interaction with antigens (46), and predict antibody developability properties (47, 48). Predicting or experimentally obtaining the native VH/VL pairing of the antibody is therefore crucial for our understanding of antibody biology and our ability to engineer these molecules (49).

Several approaches have been devised to circumvent the loss of native pairing in current Ig-seq experiments. Reddy et al. (35) assigned VH/VL pairs based on relative variable chain frequencies in VH and VL chain Ig-seq datasets. This methodology required an accompanying VL Ig-seq dataset and does not always produce antibodies with good pharmacodynamics properties, indicating that it is not always accurate (35). Researchers have also used protein expression platforms, such as recombinant cell lines or phage display, to assign VL to VH chains in a combinatorial fashion followed by experimental screening to identify productive VH/VL combinations (20, 50). Dekosky et al. (15, 51) published the first high-throughput paired VH/VL gene sequencing approach by using single-cell linkage PCR to physically join the VH and VL chains prior to Illumina sequencing. The limitation of this approach is that the current Illumina read length cannot cover the entire paired sequence, so the analysis is restricted to only CDR-H3, CDR-L3, and neighboring framework 4 and proximal positions of framework 3 of respective chains. Once sufficient paired datasets are available, these can potentially act as a reference for guiding computational pairing when VH-only Ig-seq is performed (52). Paired Ig-seq techniques currently yield smaller dataset sizes than unpaired sequencing––for instance, there were 200k sequences for the paired dataset from Dekosky et al. (15) as opposed to 40-m unpaired VH sequences in a recent study (53). The unprecedented speed and depth of Ig-seq techniques both paired and unpaired is unfortunately accompanied by high-sequencing error rates as discussed below.

The four main high-throughput sequencing platforms used to interrogate the immunoglobulin gene repertoire are Illumina, Roche 454, PacBio, and IonTorrent (39, 54–57). Earlier studies often used the Roche 454 technology as it offered greater read lengths than the Illumina methodology. In recent years, Illumina sequencing platforms are usually preferred as they have increasing read length, higher read depth, lower error rates, and lower costs per base (57, 58). Employment of unique molecular identifiers (UIDs) now permits sequencing of the entire antibody chain together with a fragment of a constant domain which holds antibody isotype information (59, 60). Unfortunately, any high-throughput Ig-seq technique suffers from significant error rates (61). Sequencing error can be introduced into Ig-seq datasets from incorrect base calling and sequencing primer artifacts, and has distinct features depending on the sequencing platform used. Error and biases can also originate from the process of preparing sequencing material including reverse transcriptase and polymerase error, amplification of nonproductive V(D)J variable domains during DNA sequencing and multiplex PCR amplification biases (62, 63). Such error may result in the overestimation of the actual number of unique clones in an Ig-seq dataset (62).

Several computational and experimental approaches have been developed to identify and remove or correct erroneous reads (58, 63), though no single-error correction strategy is currently widely used in Ig-seq repertoire analysis (30, 58). In particular, the recent application of UID to Ig-seq can help to correct errors in sequenced transcripts by generating a consensus of reads originating from the same mRNA molecule. As many studies are confined to CDR-H3 analysis, erroneous reads may also be corrected for by using a consensus CDR-H3 sequence for analysis following CDR-H3 clustering (39, 51, 64).

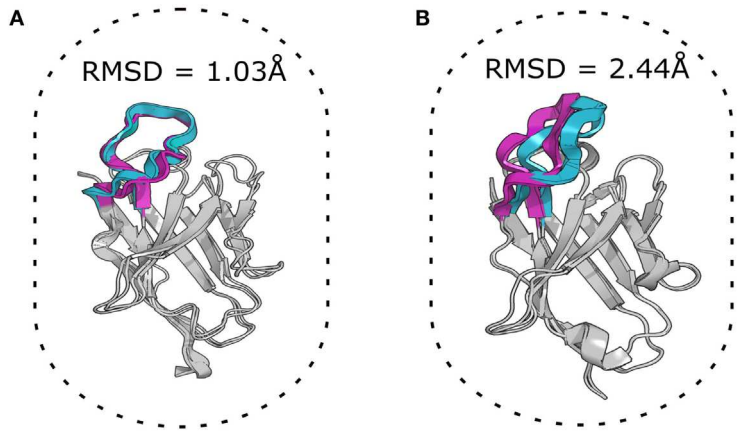
ANTIBODY STRUCTURAL PROPERTIES

The structure of an antibody is crucial in order to understand its function. Antibody–antigen recognition relies on the 3D conformation of the antibody binding site, the paratope, in relation to the cognate epitope on the antigen. In their 3D form, antibodies adopt a Y-shape conformation which can exist in monomer (IgG, IgD, and IgE), dimer (IgA) or pentamer (IgM) forms in humans (65). Several disulfide bonds help to maintain the immunoglobulin fold (**Figure 1**). One set of disulfide bonds hold the heavy constant domains together in the hinge region and another set connects the light and heavy chains (66). Intravariable domain cysteine pairs play a crucial part in shaping the antibody Fv region and artificial disruption of these bonds leads to impaired stability, folding and antigen recognition (67). These cysteines therefore have a crucial role in delineating the structural features of an antibody.

Equivalent residue positions across immunoglobulin sequences and structures can be identified by applying an antibody numbering scheme. Several numbering schemes have been developed to confer consistency and standardization on antibody sequence annotation (9, 22, 68–71). The most commonly used scheme in Ig-seq analysis is the IMGT scheme (12, 39). This numbering was built considering both structural and sequence information (9). The IMGT scheme supports symmetrical amino-acid insertions inside CDRs which ensures that structurally equivalent resides will be annotated the same regardless of CDR length. In contrast, Chothia numbering is often used by structural biologists for its simple CDR loop indel management and inherently structural focus (69, 71).

One of the principal differences between numbering schemes is how they define CDRs. Wu and Kabat (68) were the first to discover and define CDRs as portions of Fv chains that display high-sequence entropy, but as with numbering schemes, there is not a single widely adopted CDR definition and different schemes are used for legacy reasons or for specific features (such as insertion management in IMGT). The different numbering schemes define antibody CDR positions very consistently with the exception of CDR-H1 and CDR-H2 (70). Structural analysis of CDR loops has suggested that all CDRs, except for CDR-H3, adopt a restricted number of conformations, termed canonical classes (22, 72). The canonical classes link sequence patterns to a defined structure (22, 44). This enables the prediction of canonical class structure from sequence. Over the last 30 years, there have been several attempts to cluster CDR sequences/structures (22, 44, 69, 70, 72, 73). On the sequence level, the presence of certain cluster defining key residues indicates the shape the loop can adopt (22, 69, 73). Hence, some changes to the canonical CDRs can be tolerated with no explicit change to loop conformations. The different clustering methods tend to recapitulate previously found groups and find new canonical forms as a result of new data. Most algorithms incorporate CDR loops into clusters with the same number of residues (note that the number of residues varies with different CDR definitions). More recently, Nowak et al. (44) created a novel method of defining length-independent canonical classes based on findings that loops of mismatching lengths can be structurally related. This method allowed fast and accurate structural assignment of a far wider spectrum of canonical CDRs from Ig-seq datasets into fewer canonical clusters (44).

Complementarity determining region-3 of the heavy chain shows a high degree of sequence, length, and structure variation. Due to this diversity, it has so far proved impossible to classify CDR-H3 loops into canonical classes in the manner of the other CDRs. It has been proposed that CDR-H3 can be categorized into "bulged" or "extended" conformations based on the presence of asparagine at position 116 (IMGT numbering) (74, 75). However, increasing knowledge of CDR-H3 structural diversity has shown that the CDR-H3 bulged/extended configuration is difficult to predict solely from sequence (76). The relationship between sequence and structure in CDR-H3 can be important in Ig-seq as current approaches of clonotype assignment are based on CDR-H3 similarity. In this review, we define clonotypes by the presence of identical V, J genes, matching CDR-H3 lengths and CDR-H3 sequence identities greater than 85% (77). However, structural data show that CDR-H3 sequences within distinct clonotypes (sequence-dissimilar) can adopt similar 3D conformations, while those in the same clonotype (similar sequences) can adopt different 3D conformations (**Figure 2**). This suggests that the sequence alone is not a reliable indicator of similarity/difference between structures and therefore cannot



Pair	PDB	V gene	J gene	CDRH3	Sequence identity	RMSD
A	4NZU	IGHV3-30*11	IGHJ4*01	ARAPDCADADCHKGAFGY	27.7%	1.03Å
	4S1S	IGHV1-2*04	IGHJ1*01	VRTADCERDPCKGWVFPH		
B	3U7W	IGHV1-2*02	IGHJ1*01	TRGKYCTARDYYNWDFEH	88%	2.44Å
	4JDV	IGHV1-2*02	IGHJ1*01	ARGKYCTARDYYNWDFQH		

FIGURE 2 | Two aligned pairs of VH chains extracted from SAbDab, the antibody structural database (29). Complementarity determining region-3 of the heavy chain (CDR-H3) sequences in pair **(A)** belong to different CDR-H3 clonotypes but adopt very similar structural configurations with a root mean square deviation (RMSD) of ~1 Å. Pair **(B)** includes germline precursor (4JDV) and matured (3U7W) anti-gp120 antibodies (78, 79). Although CDR-H3 sequences of pair **(B)** are members of the same clonotype, the RMSD shows that their CDR-H3 shapes are structurally distinct (RMSD > 2 Å). CDR-H3 loops and their amino-acid sequences are in purple and cyan colors, mismatched amino acid are in bold. The RMSD of the backbone atom positions of proteins provides a pairwise measurement of the three-dimensional dissimilarity between two sets of coordinates where solved or predicted structures are available. Sub-Angstrom RMSD indicates structurally identical shapes, while an RMSD value greater than 2 Å for a short segment indicates structurally distinct configurations (80).

reliably indicate similar/different binding sites, functional properties and clonotype assignment.

The discrepancy between traditional clonotype assignments and native structure only illustrates how 3D information could be used to draw much more meaningful comparisons between antibodies in an Ig-seq dataset. Such comparisons should not be confined to CDR-H3 alone, but can be extended to the canonical CDRs and the entire Fv region, allowing for much more accurate grouping of functionally related antibodies.

COMPUTATIONAL TOOLS LEVERAGING ANTIBODY STRUCTURE INFORMATION

The increasing number of potential applications of antibodies as therapeutics has led to the development of computational tools which aim to streamline discovery pipelines. Some groups have already demonstrated the viability of *in silico* antibody engineering methodologies in conjunction with experimental workflows (81–84). Computational methods can be broadly divided into those that require a sequence as input and those that require a structure. Methods that require a structure as input accept experimental as well as computational models of the antibody. The large number of experimentally determined antibody structures has enabled researchers to rapidly and accurately model antibodies by leveraging homology methods (8, 85). Below we review current antibody modeling approaches and their applications.

Computational Antibody Modeling

The standard antibody modeling workflow includes four steps (**Figure 3**) (8, 86, 87). The first step is homology modeling of the VH and VL frameworks. The framework template can either be selected by sequence identity to the full-length chain (87) or to individual framework regions (8). Due to framework structure and sequence invariance, current computational tools can model framework structures very accurately (sub-Angstrom precision) (80). The second step is determining the VH/VL orientation, which can be achieved by copying the orientation angle from structures with high Fv sequence identity using VH/VL orientation methods such as AbAngle (88), analytical estimation of the angle using energy functions (89), tailored protein–protein docking (49) or structure-trained machine learning (90). Once the VH/VL orientation is set, it constrains the geometry of the binding site, allowing for the third step, which is modeling of non-H3 CDRs. At this stage, either the canonical classes are used (91) or template-based modeling such as FREAD (92) or ABGEN (93). In the final step, CDR-H3 is modeled using either homology or *ab initio* techniques (94). The resultant antibody model is refined for feasibility of dihedral angles from Ramachandran distribution, side chain orientations and side-chain clashes (89).

Homology modeling approaches can be fast at generating models if a template structure is available. Models can be created using online services: PIGSpro (86), Kotai Antibody Builder (95), and ABodyBuilder (8). Homology modeling is highly dependent on the availability of a similar template structure in

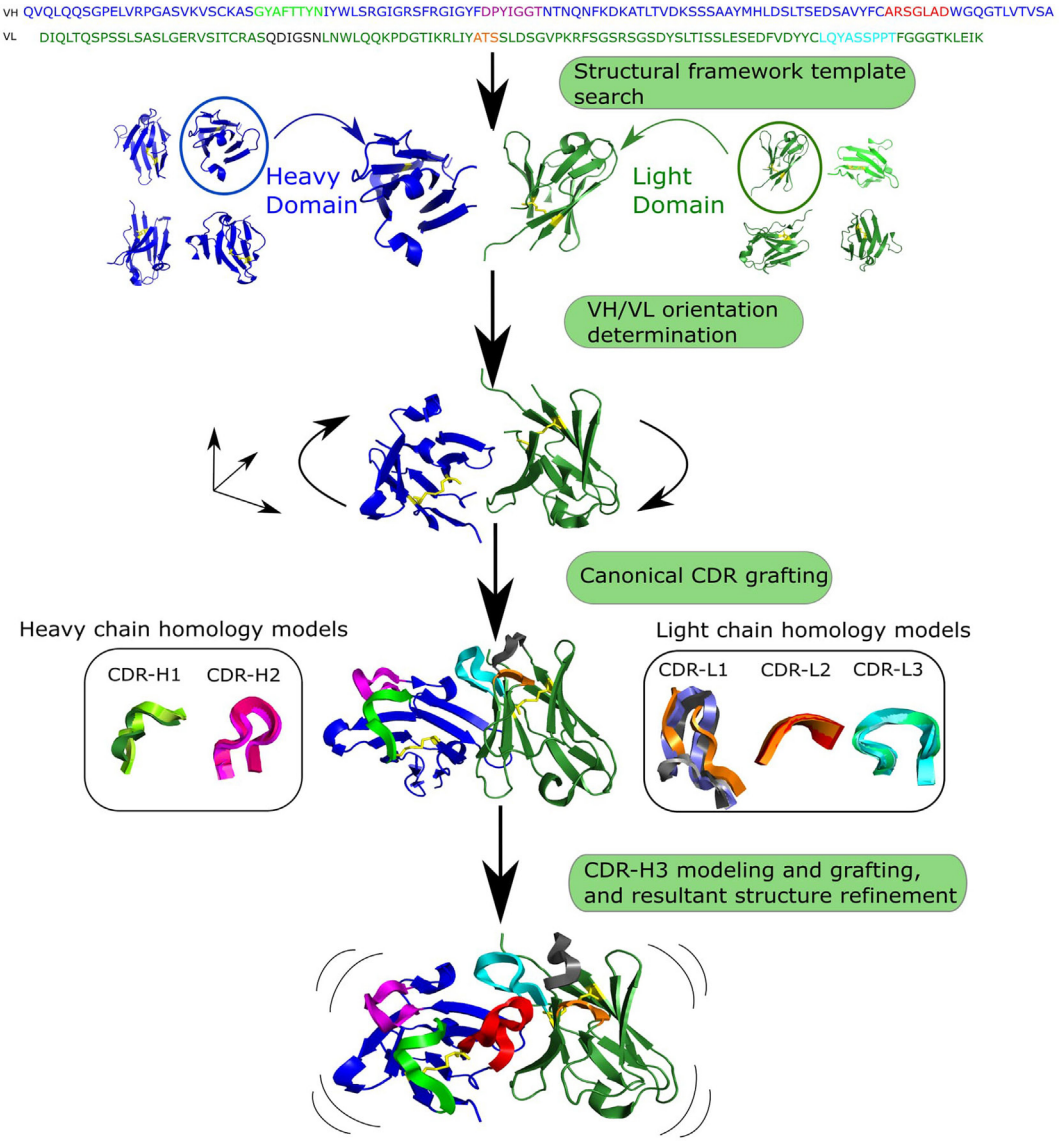


FIGURE 3 | Generalized workflow of antibody modeling. First, heavy and light chain frameworks are determined by homology modeling using templates from known structures. Next, the VH/VL orientation is calculated. The third step is modeling non-H3 complementarity determining regions (CDRs), followed by modeling and grafting of CDR-H3 onto the pre-assembled scaffold. Finally, sidechains are added to the resultant structure and it is refined.

current databases, which can be a problem for CDR-H3 where templates for longer loop length are often unavailable (94). This lack of templates is primarily due to the huge diversity of CDR-H3 shapes (96). An alternative to homology methods in such cases is *ab initio* modeling which does not rely on knowledge of already solved structures. These modeling methods create a large number of potential conformations, often referred to as decoys (97), which makes them computationally expensive compared with homology methods. *Ab initio* approaches include RosettaAntibody (98) and PLOP (99). RosettaAntibody is accessible online *via* the ROSIE (100) website, where a quick antibody modeling option is available which omits the step of intensive searching for low-energy CDR-H3 conformations. Hybrid loop modeling methodologies leverage the advantages of both modeling paradigms. For instance, Accelrys creates an initial loop model with a knowledge-based approach followed by *ab initio* loop refinement (101). More recently, a novel CDR-H3 modeling tool, Sphinx, was developed (102), inspired by the length-independent canonical CDR clustering of Nowak et al. (44). Sphinx outperformed all modeling tools on CDR-H3 structure prediction in an *ex post facto* comparison to the antibody modeling assessment (80). Despite development of different approaches, no single tool currently exists that is able to reliably model native CDR-H3 configurations. Accurate predictions of the CDR-H3 specifically and other CDRs in general are crucial to structurally characterize the antibody–antigen complex.

Performance of antibody modeling tools has been assessed in two blind studies, AMA-I and AMA-II (80, 103), where several

computational tools were benchmarked against a small number of X-ray solved but unpublished antibody crystal structures. Models of frameworks and canonical CDRs are usually accurate within 1–1.5 Å root mean square deviation (RMSD), respectively (see **Figure 2** for description of RMSD), which is very close to native structure. However, CDR-H3 prediction remains the biggest hurdle for computational antibody modeling as average accuracies for this step ranged between 2 and 3 Å RMSD, indicating a decidedly different structure to the native fold. Predictions of this quality are usually not suitable for rational design applications (80, 104).

AMA-II suggested that antibody modeling tools on average produce models of approximately similar accuracies with higher RMSD for longer loop lengths. However, the time required is radically different between homology and *ab initio* approaches (80). Homology modeling can produce a model on average in under a minute [ABodyBuilder (8)], whereas *ab initio* approaches may require up to tens of CPU hours per model [RosettaAntibody takes 482 CPU hours on average per model (100)]. To be able to use a fast homology method a suitable template is needed. Such templates are becoming more frequently available as the number of solved antibody structures increases (29). In order to model millions of sequences in a typical Ig-seq dataset, speed is crucial. Modeling at such high throughput can currently only be achieved by tools such as ABodyBuilder, which is able to generate a model within ~30 s (8). However, further increasing the rate and accuracy of antibody modeling, and developing new ways of speeding up CDR-H3 prediction, are needed if we are to structurally characterize complete Ig-seq datasets.

The accuracy and speed of some computational tools mean that thousands of sequences from Ig-seq datasets can be modeled. Such structurally annotated Ig-seq datasets allow more relevant comparisons of CDRs, binding sites and thus a more accurate grouping of molecules (**Figure 2**). The improved capacity to compare and group antibodies allows us to better visualize the antibody structure space and to investigate structural convergences of paratopes, which can be important for vaccine development (36, 37). In addition, modeled Ig-seq data can be used as input for several computational tools which annotate structure-derived antibody properties, such as therapeutic viability of the molecule (105).

Computational Prediction of Developability

Developing an antibody of high specificity and affinity against a target is only the initial step in engineering a therapeutic molecule. The resulting antibody can carry an array of risks, collectively described as developability, which includes low-expression yields, high-aggregation propensity, and off-target effects (106, 107). In the process of identifying therapeutic candidates, structurally mapped Ig-seq data can be computationally further refined for entities that pass developability criteria (45).

High-aggregation propensity is one of the most undesirable features of antibody therapeutics. Since aggregation is related to the hydrophobicity of the molecule, knowledge of structure is crucial as it allows the calculation of solvent accessible surface area. Structure-based aggregation propensity prediction tools operate by either locating surface-exposed aggregation hot spots and/or leveraging physico-chemical properties of the structure (105, 108). AGGRESCAN3D, a tool inspired by identification of hot spots in the beta amyloid peptide, distinguishes between buried, conformation engaged, and solvent-exposed aggregation prone hydrophobic patches (48). The drawback of this method was that it was not initially designed for antibodies. The Developability Index (DI) was designed for antibodies and is a structure based computational tool that quantitatively assess antibody's propensity to aggregate (105). The DI function considers the net charge of the full-length antibody and hydrophobicity of solvent-exposed sidechains of CDRs.

Such computational tools can be employed early in drug development pipelines to either isolate therapeutically viable drug candidates from the entirety of Ig-seq-derived antibody repertoire (47). Application of such structurally oriented tools requires large-scale modeling of Ig-seq datasets. Nevertheless, to date, there have not been many attempts to combine Ig-seq with structural and computational methods systematically.

COMBINING Ig-seq, STRUCTURAL, AND COMPUTATIONAL APPROACHES

Current approaches to delineate immune repertoires usually employ Ig-seq methodology only, remaining firmly within the remit of information that can be derived from sequences (31, 109, 110). The only study which has attempted to combine paired Ig-seq and structural information to characterize antibody 3D space was that of Dekosky et al. (45). Using high-throughput RosettaAntibody modeling, more than 2,000 models in naïve and antigen-experienced Ig-seq datasets were analyzed. These models helped to obtain a set of structural descriptors such as net charge, surface hydrophobicity of solvent accessible surface area for computationally determined paratopes. However, the choice of methodologies for this study imposed several limitations. Paired VH/VL data did not contain information about the full-length Fv region. Hence, all paired reads had to be completed using respective V germline gene sequences. Moreover, RosettaAntibody modeling speed only permitted the prediction of structure of 1% of the total Ig-seq dataset (2,000 sequences) in 570k CPU hours. Finally, the paired reads with CDR-H3 sequences longer than 16 amino acids were not included in the structural analysis as the modeling accuracy of such loops is currently low. This emphasizes the challenges of modeling longer CDR-H3 configurations (94, 96). Hence, novel fast and reliable CDR-H3 *ab initio* prediction as well as technologically optimized paired VH/VL gene Ig-seq are urgently needed for improved Ig-seq data modeling and interrogation.

RosettaAntibody (98) is a well-established antibody modeling tool and is able to structurally model sequence data; however, its run times make it difficult to structurally characterize the millions of sequences that are gathered during a typical Ig-seq experiment. For this reason, streamlined approaches are being developed to tackle the structural annotation of Ig-seq datasets. For instance, Nowak et al. (44) performed the structural clustering analysis of

TABLE 1 | Summary of currently available resources for computational/structural annotation of antibody sequences.

Tool type	Tool name and reference	Short tool description
ANTIBODY NUMBERING	ANARCI (113)	Variety of schemes (North, Chothia, Kabat, IMGT, AHo). Both online and command line versions are available
ANTIBODY NUMBERING	Abnum (71)	Online numbering tool that operates with Kabat and Chothia schemes
SEQUENCE ANALYSIS	IgBLAST (114)	Nucleotide and amino-acid antibody sequence analysis in IMGT and KABAT schemes
SEQUENCE ANALYSIS	IMGT/HighV-QUEST (115)	Online antibody nucleotide sequence analysis in IMGT numbering scheme
STRUCTURE DATABASE	SabDab (29)	Weekly updating database of all publically available antibody structures.
STRUCTURE/SEQUENCE DATABASE	abYsis (116)	Database of antibody structures and sequences
SEQUENCE DATABASE	DIGIT (111)	Database of antibody sequences
ANTIBODY MODELING	ABodyBuilder (8)	Homology modeling (30 s per model)
ANTIBODY MODELING	PIGSPro (86)	Homology modeling
ANTIBODY MODELING	Kotai Antibody Builder (95)	Homology modeling (90 min per model)
ANTIBODY MODELING	Accelrys (101)	Hybrid modeling (30 min per model)
ANTIBODY MODELING	RosettaAntibody (87)	*Ab initio* modeling (482 CPU hours per model)
ANTIBODY MODELING (COMMERCIAL)	Chemical Computing group (80)	Homology modeling tool combined with molecular dynamics (30 min per model)
CDR-H3 MODELING	Sphinx (102)	Length-independent hybrid modeling (30 min per model)
CDR-H3 MODELING	PLOP (99)	*Ab initio* modeling
CDR-H3 MODELING	FREAD (85)	Homology modeling (2 min per model)
PARATOPE PREDICTION	Paratome (117)	Structural consensus to identify additional antigen recognizing regions outside the CDRs
PARATOPE PREDICTION	i-Patch (118)	Statistical inference to devise a likelihood for a position to form a potential contact
PARATOPE PREDICTION	proABC (119)	Sequence-based method that leverages machine learning to predict residues that form interactions

Many of these tools have online presence and links to these are available on our website http://antibodystructure.org.

CDR-L3 of two large Ig-seq datasets: 200k paired Ig-seq sample from Dekosky et al. (15) and 9-m in-house UCB Pharma Ltd sequences as well as a database of 71k antibody sequences [DIGIT (111)]. Every CDR-L3 sequence was submitted to HMMER (112) to assign it to a length-independent cluster. This is the first instance of structurally mapping the entirety of an Ig-seq dataset. The method can be extrapolated to any non-H3 CDR to provide structural annotation of sampling of loop shapes as well as to identify yet uncharacterized loop configurations.

Structural characterization of large sequence sets can be extended to the entire Fv region. The modeling method, ABodyBuilder, was used to predict structures of 6,000 paired antibody sequences from public repositories (8). The average modeling time per 1,000 antibody sequences was 567 CPU hours compared with 285,000 CPU hours using RosettaAntibody (45). ABodyBuilder produces model accuracies that are in line with the AMA-II values (80). Using tools such as ABodyBuilder, one can perform large-scale structural modeling of Ig-seq data. Such structural characterization of Ig-seq similarity/difference would allow more accurate inter-molecule comparisons and assessment of developability. The structural software outlined in this manuscript together with other tools that are often employed in computational/structural annotation of antibody sequences is summarized in **Table 1**.

CONCLUSION

The ability to engineer better antibody-based therapeutics relies on our knowledge of the exact sequence and the 3D shape of individual molecules within the antibody repertoire. Next-generation sequencing methodologies that can yield millions of immunoglobulin gene sequences in a single sequencing run have already given insights into the steady-state and antigen-stimulated B-cell receptor repertoire (12, 32). On the other hand, low-throughput techniques such as X-ray crystallography can provide detailed information about individual antibody structures. Computational methodologies can offer a bridge between the two fields by allowing structural annotation of Ig-seq experiments (8, 44, 45). Availability of antibody structures and maturity of modeling techniques means it is now possible to perform large-scale structural characterizations of Ig-seq samples. This enriched structural content can be used to perform more precise characterization of antibodies allowing inter-antibody comparisons and grouping of structurally similar sequences (that may not be possible on the sequence level) as well as annotation of developability information. Large-scale Ig-seq datasets can also direct computational tools for targeted interrogation of antibody structural space. Statistical knowledge of the distribution of the antibody structures and sequences can offer crystallographers an idea of the common but currently unknown antibody variants. The Ig-seq and structural communities will benefit from cross-fertilization of ideas and methodologies. Together they will advance our knowledge of the antibodies in health and disease and pave the way for more advanced antibody-based therapeutics.

AUTHOR CONTRIBUTIONS

All authors contributed to the development of writing of the manuscript.

FUNDING

This work was supported by funding from Biotechnology and Biological Sciences Research Council (BBSRC) [BB/M011224/1]

and UCB Pharma Ltd awarded to AK. DK receives salary support from the NIHR Oxford Biomedical Research Centre. JT is funded by the Swiss National Science Foundation through an Ambizione-SCORE grant and has received further funding from the Olga Mayenfisch Foundation Zurich and the Bangerter-Rhyner Foundation Basel.

REFERENCES

1. Greiff V, Miho E, Menzel U, Reddy ST. Bioinformatic and statistical analysis of adaptive immune repertoires. *Trends Immunol* (2015) 36:738–49. doi:10.1016/j.it.2015.09.006
2. Vidarsson G, Dekkers G, Rispens T. IgG subclasses and allotypes: from structure to effector functions. *Front Immunol* (2014) 5:520. doi:10.3389/fimmu.2014.00520
3. Tonegawa S. Somatic generation of antibody diversity. *Nature* (1983) 302:575–81. doi:10.1038/302575a0
4. Lefranc MP. IMGT, the international ImMunoGeneTics database®. *Nucleic Acids Res* (2003) 31:307–10. doi:10.1093/nar/gkg085
5. French D, Laskov R, Scharff M. The role of somatic hypermutation in the generation of antibody diversity. *Science* (1989) 244:1152–7. doi:10.1126/science.2658060
6. Collis AVJ, Brouwer AP, Martin ACR. Analysis of the antigen combining site: correlations between length and sequence composition of the hypervariable loops and the nature of the antigen. *J Mol Biol* (2003) 325:337–54. doi:10.1016/S0022-2836(02)01222-6
7. Xu JL, Davis MM. Diversity in the CDR3 region of V H is sufficient for most antibody specificities. *Immunity* (2000) 13:37–45. doi:10.1016/S1074-7613(00)00006-6
8. Leem J, Dunbar J, Georges G, Shi J, Deane CM. ABodyBuilder: automated antibody structure prediction with data-driven accuracy estimation. *MAbs* (2016) 8:1259–68. doi:10.1080/19420862.2016.1205773
9. Lefranc M-P, Pommié C, Ruiz M, Giuducelli V, Foulquier E, Truong L, et al. IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. *Dev Comp Immunol* (2003) 27:55–77. doi:10.1016/S0145-305X(02)00039-3
10. Reichert JM. Antibodies to watch in 2017. *MAbs* (2017) 9:167–81. doi:10.1080/19420862.2016.1269580
11. Strohl WR. Current progress in innovative engineered antibodies. *Protein Cell* (2017):1–35. doi:10.1007/s13238-017-0457-8
12. Georgiou G, Ippolito GC, Beausang J, Busse CE, Wardemann H, Quake SR. The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nat Biotech* (2014) 32:158–68. doi:10.1038/nbt.2782
13. Robins H. Immunosequencing: applications of immune repertoire deep sequencing. *Curr Opin Immunol* (2013) 25:646–52. doi:10.1016/j.coi.2013.09.017
14. Lavinder JJ, Horton AP, Georgiou G, Ippolito GC. Next-generation sequencing and protein mass spectrometry for the comprehensive analysis of human cellular and serum antibody repertoires. *Curr Opin Chem Biol* (2015) 24:112–20. doi:10.1016/j.cbpa.2014.11.007
15. Dekosky BJ, Kojima T, Rodin A, Charab W, Ippolito GC, Ellington AD, et al. In-depth determination and analysis of the human paired heavy- and light-chain antibody repertoire. *Nat Med* (2014) 21:1–8. doi:10.1038/nm.3743
16. Obermeier B, Mentele R, Malotka J, Kellermann J, Kümpfel T, Wekerle H, et al. Matching of oligoclonal immunoglobulin transcriptomes and proteomes of cerebrospinal fluid in multiple sclerosis. *Nat Med* (2008) 14:688–93. doi:10.1038/nm1714
17. Lavinder JJ, Wine Y, Giesecke C, Ippolito GC, Horton AP, Lungu OI, et al. Identification and characterization of the constituent human serum antibodies elicited by vaccination. *Proc Natl Acad Sci U S A* (2014) 111:2259–64. doi:10.1073/pnas.1317793111
18. Sheynkman GM, Shortreed MR, Cesnik AJ, Smith LM. Proteogenomics: integrating next-generation sequencing and mass spectrometry to characterize human proteomic variation. *Annu Rev Anal Chem (Palo Alto Calif)* (2016) 9:521–45. doi:10.1146/annurev-anchem-071015-041722
19. Lee J, Boutz DR, Chromikova V, Joyce MG, Vollmers C, Leung K, et al. Molecular-level analysis of the serum antibody repertoire in young adults before and after seasonal influenza vaccination. *Nat Med* (2016) 22:1456–64. doi:10.1038/nm.4224
20. Chen J, Zheng Q, Hammers CM, Ellebrecht CT, Mukherjee EM, Tang HY, et al. Proteomic analysis of pemphigus autoantibodies indicates a larger, more diverse, and more dynamic repertoire than determined by B cell genetics. *Cell Rep* (2017) 18:237–47. doi:10.1016/j.celrep.2016.12.013
21. Tran NH, Rahman MZ, He L, Xin L, Shan B, Li M. Complete de novo assembly of monoclonal antibody sequences. *Sci Rep* (2016) 6:31730. doi:10.1038/srep31730
22. North B, Lehmann A, Dunbrack RL. A new clustering of antibody CDR loop conformations. *J Mol Biol* (2011) 406:228–56. doi:10.1016/j.jmb.2010.10.030
23. Li Y, Li H, Yang F, Smith-Gill SJ, Mariuzza RA. X-ray snapshots of the maturation of an antibody response to a protein antigen. *Nat Struct Mol Biol* (2003) 10:482–8. doi:10.1038/nsb930
24. Sela-Culang I, Benhnia MREI, Matho MH, Kaever T, Maybeno M, Schlossman A, et al. Using a combined computational-experimental approach to predict antibody-specific B cell epitopes. *Structure* (2014) 22:646–57. doi:10.1016/j.str.2014.02.003
25. Sircar A, Gray JJ. SnugDock: paratope structural optimization during antibody-antigen docking compensates for errors in antibody homology models. *PLoS Comput Biol* (2010) 6:e1000644. doi:10.1371/journal.pcbi.1000644
26. Krawczyk K, Liu X, Baker T, Shi J, Deane CM. Improving B-cell epitope prediction and its application to global antibody-antigen docking. *Bioinformatics* (2014) 30:2288–94. doi:10.1093/bioinformatics/btu190
27. Ecker DM, Jones SD, Levine HL. The therapeutic monoclonal antibody market. *MAbs* (2015) 7:9–14. doi:10.4161/19420862.2015.989042
28. Berman H, Henrick K, Nakamura H, Markley JL. The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res* (2007) 35:D301–3. doi:10.1093/nar/gkl971
29. Dunbar J, Krawczyk K, Leem J, Baker T, Fuchs A, Georges G, et al. SAbDab: the structural antibody database. *Nucleic Acids Res* (2014) 42:D1140–6. doi:10.1093/nar/gkt1043
30. Greiff V, Menzel U, Miho E, Weber C, Riedel R, Cook S, et al. Systems analysis reveals high genetic and antigen-driven predetermination of antibody repertoires throughout B cell development. *Cell Rep* (2017) 19:1467–78. doi:10.1016/j.celrep.2017.04.054
31. Galson JD, Pollard AJ, Trück J, Kelly DF. Studying the antibody repertoire after vaccination: practical applications. *Trends Immunol* (2014) 35:319–31. doi:10.1016/j.it.2014.04.005
32. Parola C, Neumeier D, Reddy ST. Integrating high-throughput screening and sequencing for monoclonal antibody discovery and engineering. *Immunology* (2017). doi:10.1111/imm.12838
33. Yang W, Yoon A, Lee S, Kim S, Han J, Chung J. Next-generation sequencing enables the discovery of more diverse positive clones from a phage-displayed antibody library. *Exp Mol Med* (2017) 49:e308. doi:10.1038/emm.2017.22
34. Krause JC, Tsibane T, Tumpey TM, Huffman CJ, Briney BS, Smith SA, et al. Epitope-specific human influenza antibody repertoires diversify by B cell intraclonal sequence divergence and interclonal convergence. *J Immunol* (2011) 187:3704–11. doi:10.4049/jimmunol.1101823
35. Reddy ST, Ge X, Miklos AE, Hughes RA, Kang SH, Hoi KH, et al. Monoclonal antibodies isolated without screening by analyzing the variable-gene repertoire of plasma cells. *Nat Biotechnol* (2010) 28:965–9. doi:10.1038/nbt.1673
36. Trück J, Ramasamy MN, Galson JD, Rance R, Parkhill J, Lunter G, et al. Identification of antigen-specific B cell receptor sequences using public repertoire analysis. *J Immunol* (2015) 194:252–61. doi:10.4049/jimmunol.1401405
37. Parameswaran P, Liu Y, Roskin KM, Jackson KKL, Dixit VP, Lee JY, et al. Convergent antibody signatures in human dengue. *Cell Host Microbe* (2013) 13:691–700. doi:10.1016/j.chom.2013.05.008
38. Jackson KJL, Liu Y, Roskin KM, Glanville J, Hoh RA, Seo K, et al. Human responses to influenza vaccination show seroconversion signatures and convergent antibody rearrangements. *Cell Host Microbe* (2014) 16:105–14. doi:10.1016/j.chom.2014.05.013
39. Galson JD, Trück J, Fowler A, Münz M, Cerundolo V, Pollard AJ, et al. In-depth assessment of within-individual and inter-individual variation

in the B cell receptor repertoire. *Front Immunol* (2015) 6:531. doi:10.3389/fimmu.2015.00531
40. Lowe D, Dudgeon K, Rouet R, Schofield P, Jermutus L, Christ D. Aggregation, stability, and formulation of human antibody therapeutics. *Adv Protein Chem Struct Biol* (2011) 84:41–61. doi:10.1016/B978-0-12-386483-3.00004-5
41. Tiller T, Schuster I, Deppe D, Siegers K, Strohner R, Herrmann T, et al. A fully synthetic human Fab antibody library based on fixed VH/VL framework pairings with favorable biophysical properties. *MAbs* (2013) 5:445–70. doi:10.4161/mabs.24218
42. Rouet R, Lowe D, Christ D. Stability engineering of the human antibody repertoire. *FEBS Lett* (2014) 588:269–77. doi:10.1016/j.febslet.2013.11.029
43. Krawczyk K, Dunbar J, Deane CM. Computational tools for aiding rational antibody design. In: Samish I, editor. *Methods in Molecular Biology*. Clifton, NJ: Palgrave Macmillan (2016). p. 399–416.
44. Nowak J, Baker T, Georges G, Kelm S, Klostermann S, Shi J, et al. Length-independent structural similarities enrich the antibody CDR canonical class model. *MAbs* (2016) 8:751–60. doi:10.1080/19420862.2016.1158370
45. DeKosky BJ, Lungu OI, Park D, Johnson EL, Charab W, Chrysostomou C, et al. Large-scale sequence and structural comparisons of human naive and antigen-experienced antibody repertoires. *Proc Natl Acad Sci U S A* (2016) 113:E2636–45. doi:10.1073/pnas.1525510113
46. Brenke R, Hall DR, Chuang GY, Comeau SR, Bohnuud T, Beglov D, et al. Application of asymmetric statistical potentials to antibody-protein docking. *Bioinformatics* (2012) 28:2608–14. doi:10.1093/bioinformatics/bts493
47. Kumar S, Plotnikov NV, Rouse JC, Singh SK. Biopharmaceutical informatics: supporting biologic drug development via molecular modelling and informatics. *J Pharm Pharmacol* (2017). doi:10.1111/jphp.12700
48. Zambrano R, Jamroz M, Szczasiuk A, Pujols J, Kmiecik S, Ventura S. AGGRESCAN3D (A3D): server for prediction of aggregation properties of protein structures. *Nucleic Acids Res* (2015) 43:W306–13. doi:10.1093/nar/gkv359
49. Marze NA, Lyskov S, Gray JJ. Improved prediction of antibody VL-VH orientation. *Protein Eng Des Sel* (2016) 29:409–18. doi:10.1093/protein/gzw013
50. Sato S, Beausoleil SA, Popova L, Beaudet JG, Ramenani RK, Zhang X, et al. Proteomics-directed cloning of circulating antiviral human monoclonal antibodies. *Nat Biotechnol* (2012) 30:1039–43. doi:10.1038/nbt.2406
51. DeKosky BJ, Ippolito GC, Deschner RP, Lavinder JJ, Wine Y, Rawlings BM, et al. High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. *Nat Biotechnol* (2013) 31:166–9. doi:10.1038/nbt.2492
52. Laffy JMJ, Dodev T, Macpherson JA, Townsend C, Lu HC, Dunn-Walters D, et al. Promiscuous antibodies characterised by their physico-chemical properties: from sequence to structure and back. *Prog Biophys Mol Biol* (2017) 128:47–56. doi:10.1016/j.pbiomolbio.2016.09.002
53. DeWitt WS, Lindau P, Snyder TM, Sherwood AM, Vignali M, Carlson CS, et al. A public database of memory and naive B-cell receptor sequences. *PLoS One* (2016) 11:e0160853. doi:10.1371/journal.pone.0160853
54. Rounds WH, Ligocki AJ, Levin MK, Greenberg BM, Bigwood DW, Eastman EM, et al. The antibody genetics of multiple sclerosis: comparing next-generation sequencing to sanger sequencing. *Front Neurol* (2014) 5:166. doi:10.3389/fneur.2014.00166
55. Larsen PA, Smith TPL. Application of circular consensus sequencing and network analysis to characterize the bovine IgG repertoire. *BMC Immunol* (2012) 13:52. doi:10.1186/1471-2172-13-52
56. He L, Sok D, Azadnia P, Hsueh J, Landais E, Simek M, et al. Toward a more accurate view of human B-cell repertoire by next-generation sequencing, unbiased repertoire capture and single-molecule barcoding. *Sci Rep* (2014) 4:6778. doi:10.1038/srep06778
57. Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, et al. A tale of three next generation sequencing platforms: comparison of ion torrent, pacific biosciences and illumina MiSeq sequencers. *BMC Genomics* (2012) 13:341. doi:10.1186/1471-2164-13-341
58. Friedensohn S, Khan TA, Reddy ST. Advanced methodologies in high-throughput sequencing of immune repertoires. *Trends Biotechnol* (2017) 35:203–14. doi:10.1016/j.tibtech.2016.09.010
59. Turchaninova MA, Davydov A, Britanova OV, Shugay M, Bikos V, Egorov ES, et al. High-quality full-length immunoglobulin profiling with unique molecular barcoding. *Nat Protoc* (2016) 11:1599–616. doi:10.1038/nprot.2016.093
60. Cole C, Volden R, Dharmadhikari S, Scelfo-Dalbey C, Vollmers C. Highly accurate sequencing of full-length immune repertoire amplicons using Tn5-enabled and molecular identifier––guided amplicon assembly. *J Immunol* (2016) 196:2902–7. doi:10.4049/jimmunol.1502563
61. Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, Wain J, et al. Performance comparison of benchtop high-throughput sequencing platforms. *Nat Biotechnol* (2012) 30:434–9. doi:10.1038/nbt.2198
62. Khan TA, Friedensohn S, de Vries ARG, Straszewski J, Ruscheweyh H-J, Reddy ST. Accurate and predictive antibody repertoire profiling by molecular amplification fingerprinting. *Sci Adv* (2016) 2:e1501371. doi:10.1126/sciadv.1501371
63. Shugay M, Britanova OV, Merzlyak EM, Turchaninova MA, Mamedov IZ, Tuganbaev TR, et al. Towards error-free profiling of immune repertoires. *Nat Methods* (2014) 11:653–5. doi:10.1038/nmeth.2960
64. Bokulich NA, Subramanian S, Faith JJ, Gevers D, Gordon JI, Knight R, et al. Quality-filtering vastly improves diversity estimates from illumina amplicon sequencing. *Nat Methods* (2013) 10:57–9. doi:10.1038/nmeth.2276
65. Charles A, Janeway J, Travers P, Walport M, Shlomchik MJ. The distribution and functions of immunoglobulin isotypes. *Immunobiology: The Immune System in Health and Disease*. (2001). p. 1–9. Available from: http://www.ncbi.nlm.nih.gov/books/NBK27162/
66. Wang W, Singh S, Zeng DL, King K, Nema S. Antibody structure, instability, and formulation. *J Pharm Sci* (2007) 96:1–26. doi:10.1002/jps.20727
67. Glockshuber R, Schmidt T, Plückthun A. The disulfide bonds in antibody variable domains: effects on stability, folding in vitro, and functional expression in *Escherichia coli*. *Biochemistry* (1992) 31:1270–9. doi:10.1021/bi00120a002
68. Wu TT, Kabat EA. An analysis of the sequences of the variable regions of Bence Jones proteins and myeloma light chains and their implications for antibody complementarity. *J Exp Med* (1970) 132:211–50. doi:10.1084/jem.132.2.211
69. Al-Lazikani B, Lesk AM, Chothia C. Standard conformations for the canonical structures of immunoglobulins. *J Mol Biol* (1997) 273:927–48. doi:10.1006/jmbi.1997.1354
70. Honegger A, Pluckthun A. Yet another numbering scheme for immunoglobulin variable domains: an automatic modeling and analysis tool. *J Mol Biol* (2001) 309:657–70. doi:10.1006/jmbi.2001.4662
71. Abhinandan KR, Martin ACR. Analysis and improvements to Kabat and structurally correct numbering of antibody variable domains. *Mol Immunol* (2008) 45:3832–9. doi:10.1016/j.molimm.2008.05.022
72. Chothia C, Lesk AM. Canonical structures for the hypervariable regions of immunoglobulins. *J Mol Biol* (1987) 196:901–17. doi:10.1016/0022-2836(87)90412-8
73. Martin ACR, Thornton JM. Structural families in loops of homologous proteins: automatic classification, modelling and application to antibodies. *J Mol Biol* (1996) 263:800–15. doi:10.1006/jmbi.1996.0617
74. Morea V, Tramontano A, Rustici M, Chothia C, Lesk AM. Conformations of the third hypervariable region in the VH domain of immunoglobulins. *J Mol Biol* (1998) 275:269–94. doi:10.1006/jmbi.1997.1442
75. Kuroda D, Shirai H, Kobori M, Nakamura H. Structural classification of CDR-H3 revisited: a lesson in antibody modeling. *Proteins* (2008) 73:608–20. doi:10.1002/prot.22087
76. Weitzner BD, Dunbrack RL, Gray JJ. The origin of CDR H3 structural diversity. *Structure* (2015) 23:302–11. doi:10.1016/j.str.2014.11.010
77. Chen Z, Collins AM, Wang Y, Gata BA. Clustering-based identification of clonally-related immunoglobulin gene sequence sets. *Immunome Res* (2010) 6:S4. doi:10.1186/1745-7580-6-S1-S4
78. Scharf L, West AP, Gao H, Lee T, Scheid JF, Nussenzweig MC, et al. Structural basis for HIV-1 gp120 recognition by a germ-line version of a broadly neutralizing antibody. *Proc Natl Acad Sci U S A* (2013) 110:6049–54. doi:10.1073/pnas.1303682110
79. Diskin R, Scheid JF, Marcovecchio PM, West AP, Klein F, Gao H, et al. Increasing the potency and breadth of an HIV antibody by using structure-based rational design. *Science* (2011) 334:1289–93. doi:10.1126/science.1213782
80. Teplyakov A, Luo J, Obmolova G, Malia TJ, Sweet R, Stanfield RL, et al. Antibody modeling assessment II. Structures and models. *Proteins* (2014) 82:1563–82. doi:10.1002/prot.24554

81. Clark LA, Boriack-Sjodin PA, Eldredge J, Fitch C, Friedman B, Hanf KJM, et al. Affinity enhancement of an in vivo matured therapeutic antibody using structure-based computational design. *Protein Sci* (2006) 15:949–60. doi:10.1110/ps.052030506
82. Lippow SM, Wittrup KD, Tidor B. Computational design of antibody-affinity improvement beyond in vivo maturation. *Nat Biotechnol* (2007) 25:1171–6. doi:10.1038/nbt1336
83. Thakkar S, Nanaware-Kharade N, Celikel R, Peterson EC, Varughese KI. Affinity improvement of a therapeutic antibody to methamphetamine and amphetamine through structure-based antibody engineering. *Sci Rep* (2014) 4:3673. doi:10.1038/srep03673
84. Choi Y, Hua C, Sentman CL, Ackerman ME, Bailey-Kellogg C. Antibody humanization by structure-based computational protein design. *MAbs* (2015) 7:1045–57. doi:10.1080/19420862.2015.1076600
85. Choi Y, Deane CM. FREAD revisited: accurate loop structure prediction using a database search algorithm. *Proteins* (2010) 78:1431–40. doi:10.1002/prot.22658
86. Lepore R, Olimpieri PP, Messih MA, Tramontano A. PIGSPro: prediction of immunoGlobulin structures v2. *Nucleic Acids Res* (2017) 45:W17–23. doi:10.1093/nar/gkx334
87. Weitzner BD, Jeliazkov JR, Lyskov S, Marze N, Kuroda D, Frick R, et al. Modeling and docking of antibody structures with Rosetta. *Nat Protoc* (2017) 12:401–16. doi:10.1038/nprot.2016.180
88. Dunbar J, Fuchs A, Shi J, Deane CM. ABangle: characterising the VH-VL orientation in antibodies. *Protein Eng Des Sel* (2013) 26:611–20. doi:10.1093/protein/gzt020
89. Zhu K, Day T, Warshaviak D, Murrett C, Friesner R, Pearlman D. Antibody structure determination using a combination of homology modeling, energy-based refinement, and loop prediction. *Proteins* (2014) 82:1646–55. doi:10.1002/prot.24551
90. Bujotzek A, Dunbar J, Lipsmeier F, Schäfer W, Antes I, Deane CM, et al. Prediction of VH-VL domain orientation for antibody variable domain modeling. *Proteins* (2015) 83:681–95. doi:10.1002/prot.24756
91. Marcatili P, Rosi A, Tramontano A. PIGS: automatic prediction of antibody structures. *Bioinformatics* (2008) 24:1953–4. doi:10.1093/bioinformatics/btn341
92. Deane CM, Blundell TL. CODA: a combined algorithm for predicting the structurally variable regions of protein models. *Protein Sci* (2001) 10:599–612. doi:10.1110/ps.37601
93. Mandal C, Kingery BD, Anchin JM, Subramaniam S, Linthicum DS. ABGEN: a knowledge-based automated approach for antibody structure modeling. *Nat Biotechnol* (1996) 14:323–8. doi:10.1038/nbt0396-323
94. Marks C, Deane CM. Antibody H3 structure prediction. *Comput Struct Biotechnol J* (2017) 15:222–31. doi:10.1016/j.csbj.2017.01.010
95. Yamashita K, Ikeda K, Amada K, Liang S, Tsuchiya Y, Nakamura H, et al. Kotai antibody builder: automated high-resolution structural modeling of antibodies. *Bioinformatics* (2014) 30:3279–80. doi:10.1093/bioinformatics/btu510
96. Regep C, Georges G, Shi J, Popovic B, Deane CM. The H3 loop of antibodies shows unique structural characteristics. *Proteins* (2017) 85:1311–8. doi:10.1002/prot.25291
97. Zhu K, Day T. Ab initio structure prediction of the antibody hypervariable H3 loop. *Proteins* (2013) 81:1081–9. doi:10.1002/prot.24240
98. Sircar A, Kim ET, Gray JJ. RosettaAntibody: antibody variable region homology modeling server. *Nucleic Acids Res* (2009) 37:W474–9. doi:10.1093/nar/gkp387
99. Jacobson MP, Pincus DL, Rapp CS, Day TJF, Honig B, Shaw DE, et al. A hierarchical approach to all-atom protein loop prediction. *Proteins* (2004) 55:351–67. doi:10.1002/prot.10613
100. Lyskov S, Chou FC, Conchúir SÓ, Der BS, Drew K, Kuroda D, et al. Serverification of molecular modeling applications: the Rosetta online server that includes everyone (ROSIE). *PLoS One* (2013) 8:e63906. doi:10.1371/journal.pone.0063906
101. Fasnacht M, Butenhof K, Goupil-Lamy A, Hernandez-Guzman F, Huang H, Yan L. Automated antibody structure prediction using accelrys tools: results and best practices. *Proteins* (2014) 82:1583–98. doi:10.1002/prot.24604
102. Marks C, Nowak J, Klostermann S, Georges G, Dunbar J, Shi J, et al. Sphinx: merging knowledge-based and ab initio approaches to improve protein loop prediction. *Bioinformatics* (2017) 33:1346–53. doi:10.1093/bioinformatics/btw823
103. Almagro JC, Beavers MP, Hernandez-Guzman F, Maier J, Shaulsky J, Butenhof K, et al. Antibody modeling assessment. *Proteins* (2011) 79:3050–66. doi:10.1002/prot.23130
104. Kuroda D, Shirai H, Jacobson MP, Nakamura H. Computer-aided antibody design. *Protein Eng Des Sel* (2012) 25:507–21. doi:10.1093/protein/gzs024
105. Lauer TM, Agrawal NJ, Chennamsetty N, Egodage K, Helk B, Trout BL. Developability index: a rapid in silico tool for the screening of antibody aggregation propensity. *J Pharm Sci* (2012) 101:102–15. doi:10.1002/jps.22758
106. Agrawal NJ, Kumar S, Wang X, Helk B, Singh SK, Trout BL. Aggregation in protein-based biotherapeutics: computational studies and tools to identify aggregation-prone regions. *J Pharm Sci* (2011) 100:5081–95. doi:10.1002/jps.22705
107. Trainor K, Broom A, Meiering EM. Exploring the relationships between protein sequence, structure and solubility. *Curr Opin Struct Biol* (2017) 42:136–46. doi:10.1016/j.sbi.2017.01.004
108. Tartaglia GG, Pawar AP, Campioni S, Dobson CM, Chiti F, Vendruscolo M. Prediction of aggregation-prone regions in structured proteins. *J Mol Biol* (2008) 380:425–36. doi:10.1016/j.jmb.2008.05.013
109. Vollmers C, Sit RV, Weinstein JA, Dekker CL, Quake SR. Genetic measurement of memory B-cell recall using antibody repertoire sequencing. *Proc Natl Acad Sci U S A* (2013) 110:13463–8. doi:10.1073/pnas.1312146110
110. Galson JD, Clutterbuck EA, Trück J, Ramasamy MN, Münz M, Fowler A, et al. BCR repertoire sequencing: different patterns of B-cell activation after two meningococcal vaccines. *Immunol Cell Biol* (2015) 93:885–95. doi:10.1038/icb.2015.57
111. Chailyan A, Tramontano A, Marcatili P. A database of immunoglobulins with integrated tools: DIGIT. *Nucleic Acids Res* (2012) 40:D1230–4. doi:10.1093/nar/gkr806
112. Eddy SR. Profile hidden Markov models. *Bioinformatics* (1998) 14:755–63. doi:10.1093/bioinformatics/14.9.755
113. Dunbar J, Deane CM. ANARCI: antigen receptor numbering and receptor classification. *Bioinformatics* (2015) 32:298–300. doi:10.1093/bioinformatics/btv552
114. Ye J, Ma N, Madden TL, Ostell JM. IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Res* (2013) 41:W34–40. doi:10.1093/nar/gkt382
115. Alamyar E, Duroux P, Lefranc MP, Giudicelli V. IMGT® tools for the nucleotide analysis of immunoglobulin (IG) and t cell receptor (TR) V-(D)-J repertoires, polymorphisms, and IG mutations: IMGT/V-QUEST and IMGT/HighV-QUEST for NGS. *Methods Mol Biol* (2012) 882:569–604. doi:10.1007/978-1-61779-842-9_32
116. Swindells MB, Porter CT, Couch M, Hurst J, Abhinandan KR, Nielsen JH, et al. abYsis: integrated antibody sequence and structure—management, analysis, and prediction. *J Mol Biol* (2017) 429:356–64. doi:10.1016/j.jmb.2016.08.019
117. Kunik V, Ashkenazi S, Ofran Y. Paratome: an online tool for systematic identification of antigen-binding regions in antibodies based on sequence or structure. *Nucleic Acids Res* (2012) 40:W521–4. doi:10.1093/nar/gks480
118. Krawczyk K, Baker T, Shi J, Deane CM. Antibody i-patch prediction of the antibody binding site improves rigid local antibody-antigen docking. *Protein Eng Des Sel* (2013) 26:621–9. doi:10.1093/protein/gzt043
119. Olimpieri PP, Chailyan A, Tramontano A, Marcatili P. Prediction of site-specific interactions in antibody-antigen complexes: the proABC method and server. *Bioinformatics* (2013) 29:2285–91. doi:10.1093/bioinformatics/btt369

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.